Related to other papers in this special issue	3 (p30); 4 (p40); 19 (p192); 9 (p87)
Addressing FAIR principles	F, A, I, R

FAIR Data Reuse – the Path through Data Citation

Paul Groth[1]†, Helena Cousijn[2], Tim Clark[3] & Carole Goble[4]

[1]Informatics Institute, University of Amsterdam, Amsterdam 1090 GH, The Netherlands

[2]DataCite, Welfengarten 1B, Hannover 30167, Germany

[3]Data Science Institute, University of Virginia, Charlottesville, VA 22903-1738, USA

[4]Department of Computer Science, The University of Manchester, Oxford Road, Manchester M13 9PL, UK

Keywords: FAIR data; Data citation; Research objects; Data provenance

Citation: P. Groth, H. Cousijn, T. Clark & C. Goble. FAIR data reuse – the path through data citation. Data Intelligence 2(2020), 78–86. doi: 10.1162/dint_a_00030

ABSTRACT

One of the key goals of the FAIR guiding principles is defined by its final principle – to optimize data sets for *reuse* by both humans and machines. To do so, data providers need to implement and support consistent machine readable metadata to describe their data sets. This can seem like a daunting task for data providers, whether it is determining what level of detail should be provided in the provenance metadata or figuring out what common shared vocabularies should be used. Additionally, for existing data sets it is often unclear what steps should be taken to enable maximal, appropriate reuse. *Data citation* already plays an important role in making data findable and accessible, providing persistent and unique identifiers plus metadata on over 16 million data sets. In this paper, we discuss how data citation and its underlying infrastructures, in particular associated metadata, provide an important pathway for enabling FAIR data reuse.

† Corresponding author: Paul Groth (E-mail: p.groth@uva.nl, ORCID: 0000-0003-0183-6910).

1. INTRODUCTION

Data citation has been a core part of the infrastructure in the movement toward Open Science [1]. Support for data citation was incorporated in version 1.2 of the ANSI/NISO JATS XML schema required for deposition in repositories [2] at the initiative of an expert group convened by FORCE11[①]. Major publishers and data providers have supported initiatives such as the Joint Declaration of Data Citation Principles (Data Citation Synthesis Group 2014[②]) and are rolling out support for those principles in their submission, publishing, and data archiving systems [3,4]. Support for data citation through robust data set archival and identifier generation is a common feature of many research data repositories, whether domain specific repositories like ICPSR[③] or more generic repositories like Figshare[④], Dataverse [5] and Zenodo.[⑤] DataCite alone now registers over 16 million unique identifiers (DOIs) for data sets and other non-traditional research outputs.[⑥]

It is not only that data citations are being created – they are being used [6,7], with projects underway to start measuring and exposing data reuse in the form of views, downloads, and citations of data sets [8]. Data citation has also enabled research data to begin to emerge as a first-class scholarly object, allowing the work involved to be recognized [9,10].

As others have noted (For Attribution – Developing Data Attribution and Citation Practices and Standards 2012[11]), one of the advantages of data citation is that it builds on existing scholarly practice. In Figure 1, we see an example of a data citation as it would appear in a published work. Data citations, as with any form of citation, share their existing affordances (i.e., features): they are copy-and-pastable; they provide credit through clear delineation of authorship; they give simple situatedness through a notion of a repository as a venue; they provide unique identification, and time-stamping; and they are included in the list of references. The data set that is being referred to is thus elevated to the same level as the other scholarly works (e.g., articles, books) that are being cited.

Vrba, Stian Z. & Groven, Gunleik. (2018). Stockholm reconstruction shots with lidar [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1434828

Figure 1. An example data citation.

Data citations fit – with some modifications – into existing scholarly workflows – whether it is drafting an article or building a curated list of material in a reference manager. The effort for the researcher to cite data is in some sense the same as that of a research article – figuring out the appropriate citation to use

[①] https://force11.org.
[②] https://www.force11.org/group/joint-declaration-data-citation-principles-final.
[③] https://www.icpsr.umich.edu/.
[④] http://figshare.com.
[⑤] http://zenodo.org.
[⑥] https://stats.datacite.org.

and including it in the reference list. Guidelines for data citation and their recommended format have recently been outlined by a group of publishers [3]. This ensures data citations are consistent in terms of both human and machine readability, and compatible with existing publisher practices. We still do not, for the most part, automatically generate references to data. Instead, they are included manually by researchers in the same fashion as references to research articles.

Here, we want to emphasize the, sometimes invisible, underlying capabilities available in the recommended approaches for data citation [12]. For example, a persistent identifier is required in the format of the citation. Thus, the citation is not just the string of text in an article's reference section but also enables the associated technical infrastructure to support referring to data in a unique and persistent manner. Like the data citation string itself, data citation infrastructure often also builds upon existing scholarly infrastructure. However, it expands this infrastructure to enable new functionality that provides a strong foundation for not just referring to data, but injecting it into the scholarly ecosystem and making it more reusable.

The aim of this paper, is to introduce an exemplar data citation infrastructure as implemented by DataCite, a global non-profit organization that provides persistent identifiers (DOIs [13]) for research data and other research outputs, and to show how its capabilities may be used to enhance the reusability of data. We note that data citation infrastructures such as identifiers.org and ARKS also support many of the capabilities we will discuss [14]. The important point is to illustrate what these capabilities are. We hope that this can serve as guide for data providers to use the capabilities of these infrastructures more completely. Just as data citation builds on existing scholarly practice, so too can the move toward producing more Findable, Accessible, Interoperable and Reusable Data [15] built on the success of data citation.

The rest of this article is organized in four sections. First, we begin by introducing the data citation infrastructure, followed by a discussion of the role of metadata. We then address the use of data citation infrastructure for both expressing data provenance and contextualizing data for reuse. Finally, we touch on the need for the grounding of data citation in the scientific social ecosystem through the scholarly literature.

2. UNDERSTANDING DATA CITATION INFRASTRUCTURE

As Figure 1 shows, after the authors, title, and archival repository of the data set, the citation ends with a persistent identifier. A persistent identifier[7] is a long-lasting reference to an object and usually directs to a landing page with information about the underlying object. The main idea is that over time the location referred to by the identifier will either still exist, or will need to redirect to a new location for the object, or will state that the object is no longer available. Often DOIs, as in the case of DataCite citations, are used for these persistent identifiers. Other identifiers like ARKs or CURIEs can also be used [14]. In each case, there is an intermediary who is responsible for creating (i.e., registering) these identifiers. At an institutional level, intermediaries such as DataCite (in the case of DataCite DOIs) or California Digital

[7] https://www.ands.org.au/guides/persistent-identifiers-expert.

Library (in the case of ARKs) provide social guarantees about the longevity of these identifiers, and work with institutions to ensure that these are redirected as the underlying institutional infrastructures change. In our example case (Figure 1), the DOI redirects the user to a Landing Page URI at the Zenodo data repository. If for some reason Zenodo changes its URL scheme or ceases to exist, the DOI can be redirected to another location, thus maintaining access to the data.

This intermediation is a critical component of such systems, not just from a social or longevity perspective, but also from a technical perspective. When organizations register persistent identifiers for their data with DataCite, they also deposit associated metadata which is then hosted by the intermediary. This is done following a metadata schema[⑧] so that metadata terms are clearly and interoperably defined, and consistently provided. For example, the intermediary can guarantee that information about the author of the metadata is always accessible using the same schema property. Thus, while data citation is often seen as largely addressing the "Findable" and "Accessible" components of the FAIR principles, it is worth emphasizing their role in "Interoperability" and "Reusability".

3. METADATA AND DATA CITATION INFRASTRUCTURE

Data citation intermediaries provide a convenient home for the addition of metadata that should be available for any digital object. What does this mean in practice? Figure 2 shows example metadata from the data citation above, retrieved from our example citation. Figure 2a shows the redirection, the title and author names. However, Figure 2b shows the beginnings of the power of data citation metadata. Here, what we see is additional metadata beyond that in the citation itself, enabling the provision of information useful in determining the reusability of data. In Figure 2b, we see that the data have a well-defined Creative Commons license, is open access, was funded by the European Commission, and has a prior version.

[⑧] https://schema.datacite.org/.

(a)

```
"id": "https://doi.org/10.5281/zenodo.1434827",
  "doi": "10.5281/ZENODO.1434827",
 "url": "https://zenodo.org/record/1434827",
  "creators": [
              {
              "name": "Vrba, Stian Z.",
              "nameType": "Personal",
              "givenName": "Stian Z.",
              "familyName": "Vrba",
              "affiliation": "Quine AS"
              },
              {
              "name": "Groven, Gunleik",
              "nameType": "Personal",
              "givenName": "Gunleik",
              "familyName": "Groven"
              }
  ],
  "titles": [
              {
              "title": "Stockholm Reconstruction Shots
With Lidar"
              }
```

(b)

```
"rightsList": [
               {
               "rights": "Creative Commons Attribution 4.0",
               "rightsUri": "https://creativecommons.org/licenses/
by/4.0"
               },
               {
               "rights": "Open Access",
               "rightsUri": "info:eu-repo/semantics/openAccess"
               }
 ],
 "fundingReferences": [
               {
               "awardUri":     "info:eu-repo/grantAgreement/EC/
H2020/731970/",
               "awardTitle": "Live Action Data Input and Output",
               "funderName": "European Commission",
               "awardNumber": "731970",
               "funderIdentifier": "https://doi.org/10.13039/50110
0000780",
               "funderIdentifierType": "Crossref Funder ID"
               }
 ],
 "relatedIdentifiers": [
               {
               "relationType": "HasVersion",
               "relatedIdentifier": "10.5281/zenodo.1434828",
               "relatedIdentifierType": "DOI"
               }
```

Figure 2. Example DataCite metadata[9].

Therefore, intermediation provides two benefits. First, it offers a convenient location to provide additional metadata. Secondly, it drives the standardization of core metadata on data sets and other digital research objects. For example, DataCite was able to easily provide schema.org formatted metadata for the data sets registered with it, improving availability of data set information for search engines to index [16]. The DataCite Metadata Schema (DataCite Metadata Working Group 2019[10]) provides a wide variety of properties ranging from denoting the type of contribution someone made (e.g., Project Leader, Editor) to the specific geolocation to which data are related. This is just a tiny fraction of the available descriptor space. Thus, hosting the metadata necessary for FAIR data, data citation intermediaries provide permanent and easy access to consistent metadata. In the next section, we discuss a set of metadata properties already available for use that would greatly enhance data reuse.

[9] https://api.datacite.org/dois/application/vnd.datacite.datacite+json/10.5281/zenodo.1434827.

[10] DataCite Metadata Working Group. 2019. "DataCite Metadata Schema Documentation for the Publication and Citation of Research Data v4.2." DataCite. https://schema.datacite.org/meta/kernel-4.2/index.html.

4. PLACING DATA IN THE GLOBAL PROVENANCE GRAPH

A critical part of understanding whether data can be reused is to understand how they fit in a larger context. This includes understanding how data were produced – their provenance [17,18], unambiguous description of the concepts under consideration, and relationships to other sources. The DataCite metadata schema has introduced the notion of relation types. These 32 types allow the expression of many types of relations including for example, that a data set is derived from another data set (isDerivedFrom); that a data set is a new version of a data set (isNewVersionOf); that a data set is documented by a particular piece of documentation (isDocumentedBy); or that a data set is created using a piece of software (isCompiledBy). By asserting these links, a data set provider can express the provenance of the data [19].

Importantly, it is not just that the data citation allows for the expression of links between data but also links into the existing literature citation graph. Thus, data can be contextualized not only by their relation to data and software but by their relation to the scholarly discourse.

Beyond provenance and the literature context, it is also possible to express the actual entities a data item is about through the use of specific subject identifiers. Here, the emergence of Wikidata is of interest in the stabilization and coalescence of conceptual terminology because of the diversity of scholarly communities with established languages and vocabularies. Wikidata provides a global space for referring to common entities and concepts in a language independent fashion [20]. It provides large numbers of definitions and multilingual links and allows data to be contextualized within a large global knowledge base of facts. Wikidata is increasingly being used to provide a common linking point across research databases [21, 22]. Thus, by providing links to this common space through the subject identifier metadata property available, the specific subjects can be defined in this common space. This, for example, could enable one to find all the data sets about a particular gene or protein.

In all these cases, the links are expressed through globally unique persistent identifiers. The provenance graph and other context information thus become part of the global scientific record [23].

While context is critical for reuse, even more interesting is to be able to potentially regenerate or build upon a data set by having access to its entire experimental context in the form of a Research Object [24], which we now discuss.

5. BUILDING RESEARCH OBJECTS USING DATA CITATION RECORDS

A Research Object[⑪] is a bundle of all the artefacts associated with an investigation or piece of research into one whole or package that can itself be cited [24]. This may be done by packing a set of elements into a container (e.g., a zip file or a BagIt file [25] with a manifest file that describes the contents. The manifest metadata describes the relationship between elements within the bundle. Data citation metadata provides

[⑪] http://www.researchobject.org/overview/.

another possible route to bundling these elements together and exposing the elements of a research object together in an accessible fashion.

The following is a sketch of how this could be done. First, one would create a research object or research object stub. Using DataCite Metadata terminology this would be a Collection. By using the aforementioned relationship types, one can express the relationship to the software, workflows, data, documentation and papers that are all members of the collection. Importantly, these relationships are already expressible using DataCite metadata. The collection then defines the distinct research object package while not necessarily needing to encapsulate all parts and instead holding references to the constituent parts. Additionally, because data citation supports versioning natively, one can express accurately the notion of the true contents of a research object.

Such packaging is crucial for reuse as data themselves do not stand alone, it is by their nature contextualized by both the computational environment in which they can be generated and used, and their broader social embedding [26].

6. REUSE AND THE IMPORTANCE OF THE HUMAN

As discussed above, data citation provides a critical component often lost in the discourse around machine reusability, which is the link to the human. While the goal to promote the ability for machines to understand data is an exciting one, it is poor scholarly practice to reuse data without understanding their original context and conditions of creation [27, 28]. In addition, providing all the necessary elements to generate completely machine reusable data may be too resource intensive, outside of the most high-value data [29]. In the end it is the responsibility of the researchers to understand the nature of data, and the appropriate conditions for data reuse, employing the associated metadata and literature artefacts we provide them. By linking data to metadata and to the literature or other human readable documentation, data citation provides a critical outlet to facilitate reusability.

7. CONCLUSION: DATA CITATION INFRASTRUCTURE AS SCAFFOLDING

In this short article, our aim was to point out the powerful features that are already available in data citation infrastructures that make it amenable to supporting reusable data. However, it is just that – support. Data citation infrastructure does not provide the metadata, it provides a uniform place to deposit and access it. Given this, we think there are some steps that tool builders and data repositories should implement to facilitate reuse of data based on this infrastructure:

1). Make it easy to express the relationships supported by the data citation infrastructure and publish those in the associated metadata repository.
2). Promote the use of subject identifiers, in particular from Wikidata.
3). Develop tools that contextualize data within the larger scholarly ecosystem.
4). Ensure both data creators and users are aware of the possibilities to add and use metadata.

We encourage data providers, data hosts, and the entire FAIR data community to consider how we can use this already available scaffolding to expose the elements needed to make reusable data.

AUTHOR CONTRIBUTIONS

P. Groth (p.groth@uva.nl) conceptualized and wrote the first draft of the paper. T. Clark (twc8q@virginia.edu), H. Cousijn (hcousijn@datacite.org) and C. Goble (carole.goble@manchester.ac.uk) clarified the ideas and concepts in the paper. All authors edited and reviewed the final version of the article.

ACKNOWLEDGEMENTS

This work was partially supported by Horizon 2020, INFRADEV-4-2014-2015, 654248, CORBEL, Coordinated Research Infrastructures Building Enduring Life-science services.

REFERENCES

[1] G. Silvello. Theory and practice of data citation. Journal of the Association for Information Science and Technology 69(1)(2018), 6–20. doi: 10.1002/asi.23917.

[2] D. Mietchen, J. McEntyre, J. Beck, C. Maloney & Force11 Data Citation Implementation Group. Adapting JATS to support data citation. In: Journal Article Tag Suite Conference (JATS-Con) Proceedings 2015 [Internet], National Center for Biotechnology Information (US). Available at: https://www.ncbi.nlm.nih.gov/books/NBK280240/.

[3] H. Cousijn, A. Kenall, E. Ganley, M. Harrison, D. Kernohan, T. Lemberger, ... & T. Clark. A data citation roadmap for scientific publishers. Scientific Data 5(2018), Article No. 180259.

[4] M. Fenner, M. Crosas, J.S. Grethe, D. Kennedy, H. Hermjakob, P. Rocca-Serra, ... & T. Clark. A data citation roadmap for scholarly data repositories. Scientific Data, 6(2019), Article No. 28. doi: 10.1038/s41597-019-0031-8.

[5] G.King. An introduction to the Dataverse Network as an infrastructure for data sharing. Sociological Methods and Research 32(2)(2007), 173–199. doi: 10.1177/0049124107306660.

[6] M.S. Mayernik & K.E. Maull. Assessing the uptake of persistent identifiers by research infrastructure users ed. Ruslan Kalendar. PLOS ONE 12(4)(2017), e0175418. doi: 10.1371/journal.pone.0175418.

[7] N. Robinson-García, E. Jiménez-Contreras & D. Torres-Salinas. Analyzing data citation practices using the data citation index. Journal of the Association for Information Science and Technology 67(12)(2016), 2964–75. doi: 10.1002/asi.23529.

[8] H. Cousijn, P. Feeney, D. Lowenberg, E. Presani & N. Simons. Bringing citations and usage metrics together to make data count. Data Science Journal 18(1)(2019), 9. doi: 10.5334/dsj-2019-009.

[9] M. Altman & M. Crosas. The Evolution of Data Citation: From Principles to Implementation. IASSIST Quarterly 37(1)(2013), 62–70. doi: 10.29173/iq504.

[10] B.E. Bierer, M. Crosas, H.H. Pierce. 2017. Data authorship as an incentive to data sharing. The New England Journal of Medicine 376(17)(2017), 1684-1687. doi: 10.1056/NEJMsb1616595.

[11] For Attribution — Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop. 2012. Washington, D.C.: National Academies Press. Available at: http://www.nap.edu/catalog/13564.

[12] Joan Starr, Eleni Castro, Mercè Crosas, Michel Dumontier & T. Clark Achieving human and machine accessibility of cited data in scholarly publications. PeerJ Computer Science 1, e1. doi: 10.7717/peerj-cs.1.

[13] International DOI Foundation. The DOI Handbook. (2012). doi: 10.1000/186.

[14] S.M. Wimalaratne, N. Juty, J. Kunze, G. Janée, J.A. McMurry, N. Beard, ... & Tim Clark. Uniform resolution of compact identifiers for biomedical data. Scientific Data 5(2018), Article No. 180029. doi: 10.1038/sdata.2018.29.

[15] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, … & B. Mons. The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3(2016), Article No. 160018. doi: 10.1038/sdata.2016.18.

[16] M. Fenner. 2017. Using Schema.org for DOI Registration. doi: 10.5438/0000-00cc.

[17] M. Herschel, R. Diestelkämper & H.B. Lahmar. A survey on provenance: What for? What form? What from? The VLDB Journal 26(6)(2017), 881–906. doi: 10.1007/s00778-017-0486-1.

[18] L. Moreau & P. Groth. Provenance: An introduction to PROV. San Rafael, CA: Morgan & Claypool Publishers, 2013.

[19] M. Fenner. Exposing DOI metadata provenance (Version 1.0). (April 10, 2019). doi: 10.5438/wy92-xj57.

[20] D. Vrandečić & M. Krötzsch. Wikidata: A free collaborative knowledgebase. Communications of the ACM 57(10)(2014), 78–85. doi: 10.1145/2629489.

[21] Burgstaller-Muehlbacher, Sebastian et al. 2016. Wikidata as a semantic framework for the gene Wiki initiative. Database 2016 baw015.

[22] F.A. Nielsen, D. Mietchen & E. Willighagen. Scholia, scientometrics and Wikidata. In: E. Blomqvist, K. Hose, H. Paulheim et al. (eds.) The Semantic Web: ESWC 2017 Satellite Events. ESWC 2017. Cham, Switzerland: Springer, 2017, pp. 237–259. doi: 10.1007/978-3-319-70407-4_36.

[23] M. Fenner & A. Aryani. Introducing the PID Graph. Available at: https://blog.datacite.org/introducing-the-pid-graph/ (April 16, 2019).

[24] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J., Bhagat, … & C. Goble. Why linked data is not enough for scientists. Future Generation Computer Systems, 29(2)(2013), 599–611. doi: 10.1016/j.future.2011.08.004.

[25] J. Kunze, J. Littman, E. Madden, J. Scancella & C. Adams. The BagIt file packaging format (V1.0). (October 2018). doi: 10.17487/rfc8493.

[26] J. Vertesi & P. Dourish. The value of data: Considering the context of production in data economies. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW '11), 2011, pp. 533–542. doi: 10.1145/1958824.1958906.

[27] C.L. Borgman. Big data, little data, no data: Scholarship in the networked world. Cambridge MA: MIT Press, 2015.

[28] I. Pasquetto, B. Randles & C. Borgman. 2017. On the reuse of scientific data. Data Science Journal, 16(2017), 8. Available at: http://datascience.codata.org/articles/10.5334/dsj-2017-008/.

[29] C.L. Borgman. 2012. The conundrum of sharing research data. Journal of the American Society for Information Science and Technology 63(6)(2012), 1059–78. doi: 10.1002/asi.22634.